

Coloured graphlet profiles as a predictor of career length in scientific co-authorship networks

Oliver Blanthorn
School of Computer Science
The University of Manchester
Manchester, United Kingdom
oliver.blanthorn@manchester.ac.uk

Eva Navarro-López*
School of Computer Science
The University of Manchester
Manchester, United Kingdom
eva.navarro@manchester.ac.uk

Abstract—Graphlets, or induced motifs, have long been used to find important medium-scale structures in directed networks. We present a method using the composition of coloured graphlets in ego-networks to characterise nodes. We give an example application using our technique to predict the numbers of years researchers are active from their collaboration networks, and compare our success with simpler metrics; particularly, we find that the use of coloured graphlets improves predictive performance compared to colour-blind graphlets; that 4-star graphlets centred on an author are predictors of a long career, and that this effect is not degenerate to centralities.

Keywords—graphlets, motifs, colour, collaboration, co-authorship, career length, science of science

I. INTRODUCTION

A complex network is a way of describing the interactions of a complex system [1]; it is a set of interacting objects (nodes), and links between those nodes, representing interactions. For example, in a co-authorship network, the nodes are individual authors, and the links between the authors represent collaborations. Many complex networks are evolving or dynamic, just like the complex systems they represent. This paper examines our ability to predict the length of time a node will be active in a network from its local network structure. We give an example application in the form of prediction of researchers' career lengths from a dynamic co-authorship network.

Graphlets, the small induced subgraphs that form the building blocks of networks, have long been used to characterise the local structure of complex networks [2], [3]. More recently, it has been suggested that graphlets may be effective for the characterisation of single nodes through their ego-networks, the induced subgraphs containing a node and all of its neighbours a certain distance away [4]. We present some extensions to this technique, particularly considering undirected ego-networks coloured to distinguish the ego, and compare it to other methods of characterisation of nodes, namely the standard centralities such as node degree, betweenness centrality, and the local clustering co-efficient.

We apply this node characterisation technique in a specific application: predicting the “lifetime” of researchers from the shape of their collaborations. This application is particularly important as the research lifetime of researchers, that is, the number of years they are active publishing research, follows a

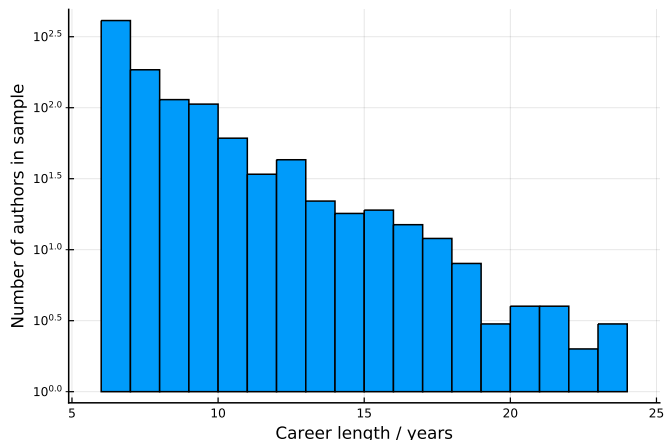


Fig. 1. A log-linear histogram of showing the distribution of researcher lifetimes in the subsample of the Arnetminer co-authorship data set [8] used throughout this paper. The lowest career length is 5 and the highest career length is 23. Both of these artefacts are due to choices we have made to ensure we do not have an incomplete picture of any careers.

power-law, so the overwhelming majority of researchers have short research lifetimes, as shown in Fig. 1, and found in other research [5]. It has been suggested that this distribution is due to the voluntary career changes of PhD students [6], but this would imply a bi-modal distribution; instead, the power-law has a roughly constant exponent, suggesting that a constant proportion of researchers leave academia each year. 53% of all UK workers do not wish to change career [7], suggesting that a large proportion of career changes away from academia are involuntary. Our motivation is to find information that could help early career researchers extend their careers in academia, as many of them clearly wish to do so. The shape of a collaboration network is relatively easy to change as one chooses who to collaborate with. Our work aims to uncover whether particular shapes of research collaboration predict career longevity, and to explore the importance of colour, or nodes which fall into discrete categories, in ego-networks.

The rest of this paper is organised as follows: Sect. II gives an overview of the terminology and previous work in this area needed to understand our work; Sect. III explains our methodological contributions; Sect. IV demonstrates the use

of our technique with a real-world application; and Sect. V summarises our results and possible directions for future work.

II. BACKGROUND

In this section, we give an overview of some of the mathematical and computational tools that this work is based on: broadly, graphlets, co-authorship networks, ego-networks and colour.

A. Motifs and graphlets

Motifs are characteristic subgraphs, tallied by automorphism group—that is, graphs whose nodes can be relabelled to produce an identical adjacency matrix—of a graph. The “characteristic” refers to subgraphs that appear more often than one would expect from a null model [2]. Graphlets are a similar concept, originally with the distinction that the subgraphs must be induced and that no significance test or normalisation is required. Many authors now use motifs to refer to induced subgraphs, i.e., subgraphs defined by a subset of nodes and the edges between those nodes.

Graphlets have been used to characterise nodes through their ego-networks [4], [9] through choosing a few graphlets to track over time, and counting the frequencies of their appearances. The result of graphlet enumeration is a graphlet profile: a mapping from graphlets to frequencies. Uncoloured graphlets in co-authorship networks have been used to predict success in terms of citation count [10]. It was found that “box motifs”, i.e., 4-cycles, were beneficial.

There are many methods for counting graphlets and motifs [11]. Computation time is not a limiting factor for our applications as the ego-networks usually only have tens of nodes, so we have used Kavosh, an exact method [12]. Approximate methods may have better computational performance.

B. Co-authorship networks

A co-authorship network is a type of collaboration network whose nodes are authors and whose links signify co-authorship of a work. One of the main challenges with real-world co-authorship networks is disambiguation: the ability to identify when two authors with the same name are different people. data sets exist where this procedure has been completed [8].

As with other evolving networks made from data where events are recorded as instantaneous but in reality have some unknown duration, a window size is chosen for which the events are extended. For example, a 3-year window size means that a co-authorship registered in a certain year is extended to also include the two years before it; many collaboration network data sets effectively set the window size to infinity and consider the network as if it were static.

For a detailed overview, see [13]. Some work has been conducted with graphlets and motifs on co-authorship networks [14], [15], but coloured graphlets have been neglected; some authors have considered the link between node centralities, career length and citation counts [16].

C. Ego-networks and colour

An n -ego-network is a subgraph induced by the node of interest, the ego, and all of its neighbours a certain distance n or less from the ego; these neighbours are called the alters [17]. Coloured graphs are graphs whose nodes fall into discrete categories, which can be applied trivially to ego-networks by setting the ego to be one colour, and the alters to another. In our application, the ego is the author whose lifetime we predict, and the alters are their co-authors, their co-authors’ co-authors, and so on.

III. METHODS

We predict the lifetimes of nodes in an evolving complex network using a simple linear regression model that combines all of our features. This allows us to disregard degenerate features. Our features consist of graphlet profiles and the standard node centralities.

Our software is written in Julia [18] using the package LightGraphs.jl [19]. LightGraphs.jl is up to $100\times$ faster than NetworkX and is comparable to graph-tool; Julia allows for concise performant code. Plots.jl was used for plotting, and GLM.jl was used for prediction.

A summary of our technique follows:

- 1) Clean data—ensure that we have a complete lifecycle for all nodes considered,
- 2) choose a subsample of these nodes to limit computation time,
- 3) calculate lifetimes of the chosen nodes,
- 4) extract the 3-ego-networks of each node at a specific early stage in their lifecycle,
- 5) extract features: calculate centralities and graphlet profiles on these ego-networks,
- 6) calculate the linear correlation of features with lifetime,
- 7) check the pairwise correlation between features to check for possible degeneracy,
- 8) and finally, perform multiple linear regression to determine whether correlations are better explained by other features.

A. Ego-networks and coloured graphlets

We chose to consider the 3-ego-networks of each node, where 3 corresponds to the maximum distance from the node of interest. Three was chosen as it was found that any more than three did not improve predictions and led to somewhat increased computational complexity, considerably so in the worst cases; fewer than three would have meant that some 4-graphlets were underrepresented.

The ego in each ego-network is given a colour distinct from all of the other nodes, used in the subgraph matching. This allows us to understand the importance of an ego within a graphlet, if it is there at all.

Each graphlet is given a label with the convention that the number before the decimal place is the number of nodes in the graphlet, the number after the decimal place labels the monochromatic automorphism group in ascending order of the Shannon entropy of their degree distributions (so less-uniform

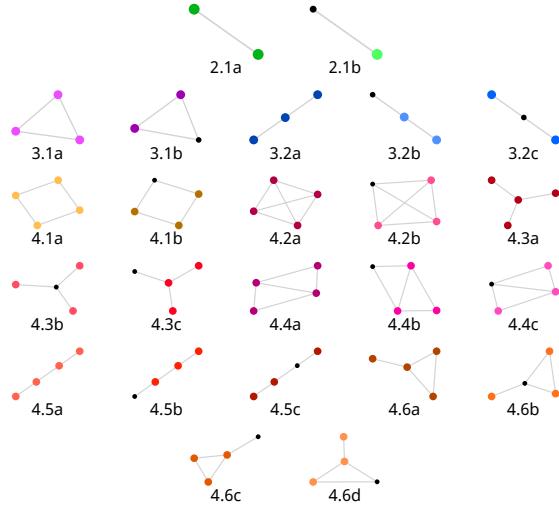


Fig. 2. All possible undirected graphlets with at most a single differently-coloured node (i.e., the possible graphlets within ego-networks where the ego is one colour and the alters are another colour) up to size 4. The smaller black node corresponds to the ego, if it exists, and the coloured nodes are the alters. The graphlets are labelled as described in Sect. III-A; the colours are chosen such that similar graphlets have similar colours. The same colours are used throughout this work.

distributions have a higher label), and the character signifies colour differences between graphlets within the same monochromatic automorphism group, as shown in Fig. 2.

B. Graphlet profiles

We use our own partial implementation of Kavosh [12] to enumerate the graphlets and create the graphlet profiles of the ego-networks. Reimplementation allowed us to mitigate the integer overflows that the reference Java implementation was found to suffer from, and let us trivially extend Kavosh to be aware of coloured subgraphs with Nauty [20].

We normalise each graphlet profile of graphlets with n nodes, g_n , by the total number of induced subgraphs of size n in the network,

$$\hat{g}_{ni} = \frac{g_{ni}}{\sum_i g_{ni}}, \quad (1)$$

where each ni corresponds to a specific automorphism group i of size n . We feel that this is sufficient compared to the much more computationally expensive significance profiles approach of using the ratio of the counts to an ensemble of configuration models [2] as we are later comparing these profiles with many other profiles: we are not concerned with how they compare to configuration models: we want to know how they differ from other networks in same domain.

After normalisation, we combine the graphlet profiles for $n = 2, 3, 4$ in a single array, g . Larger graphlets are not used due to their increased computational complexity and our experimental findings that they did not improve predictions, in agreement with previous studies [4].

We produce coloured and uncoloured graphlet profiles and compare them; we refer to them as the coloured and uncoloured treatments. Monochromatic graphlets can appear in

TABLE I
FEATURES USED TO CHARACTERISE EGOS IN EGO-NETWORKS

Feature	Description	Range
Betweenness ^a	Prop. of shortest paths passing through ego	[0, 1]
Degree ^a	Number of edges attached to ego	[0, ∞)
Closeness ^a	Reciprocal of distance of ego to alters	(0, 1]
Clustering	Ratio of triangles present to possible triangles	[0, 1]
Ego-clustering	Clustering of 1-ego-network of ego	[0, 1]
Edges	Number of edges in ego-network	[0, ∞)
Graphlet profile	Normalised graphlet counts as described	[0, 1] ^d

^a Commonly known with the suffix “centrality”, omitted here for brevity

both types of profile but we refer to them with a trailing “a” in the coloured treatment to signify that they represent the complete absence of the ego.

C. Other centralities considered

Table I summarises all of the features we consider in this work. We refer to all of the features that are not graphlet profiles as the “standard centralities”, but not all of the features are centralities in the traditional sense as they characterise graphs rather than nodes. However, because we are using them to describe the ego-network of specific nodes, they can in some sense be considered to characterise those nodes, so we take the liberty of referring to them as centralities.

D. Estimation of confidence intervals

We estimate errors via a “bootstrap method”: the standard error is calculated for the correlation with lifetime for 10 subsamples of the graphlet profiles. This error is then scaled by the square root of the number of subsamples. The error bars plotted correspond to twice the standard error and are therefore approximations of the 95% confidence interval, assuming a normal distribution.

E. Multiple linear regression

Multiple linear regression is a simple method for investigating the relationship between many continuous so-called explanatory variables x_i , and a single dependent variable, y , with

$$y = \sum_i \beta_i x_i,$$

where β_i are co-efficients to be determined by the least-squares method [21]. It is also common to report the β_{reduced} co-efficients, which are the co-efficients multiplied by the standard deviation of each explanatory variable, to ease interpretation of the co-efficients. Each reduced co-efficient corresponds to the change one in the dependent variable one could reasonably expect as a result of increasing the corresponding explanatory variable from a figure representative of the samples with a low amount of that variable, to a figure representative of samples with a high amount of that variable.

In this work, the explanatory variables are the features enumerated in Table I and the dependent variable in our application is the research lifetime of the authors.

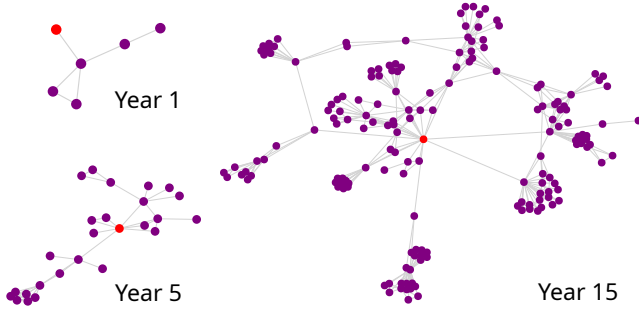


Fig. 3. The 3-ego-networks, anti-clockwise from top-left, of one researcher’s first, fifth and fifteenth year in their career as a researcher. The ego is shown in red; the alters are shown in purple.

It is easy to perform statistical tests on this model to estimate the probability that the co-efficients and their standard errors found could be due to random chance rather than a real link. In our results, we display the t value—the ratio of the co-efficient to the standard error—and the probability that one would receive a t value of that size or larger due to random fluctuations; generally, one interprets this as the chance that the true value of the co-efficient is zero. This probability is referred to variously as $Pr(> |t|)$ and the p -value.

This p -value is only valid in the context of a single, isolated, hypothesis test (where the null hypothesis is that the co-efficient is zero). The large number of possible graphlets combined with the other centralities means that we will be testing many hypotheses simultaneously, greatly increasing the possibility of false positives. We therefore apply a Bonferroni correction to guard against this—we check the p -values against a corrected p -value of $\frac{p}{n}$, where p is the desired p -value, and n is the number of hypotheses tested [22]. Since we have 30 features, our corrected p -value corresponding to a 95% confidence level is approximately 0.0017.

IV. APPLICATION TO TEMPORAL CO-AUTHORSHIP NETWORK

Here we present an application of our technique to the prediction of research lifetimes from a temporal co-authorship network.

A. Experimental design

The data set chosen was Arnetminer’s disambiguated temporal collaboration network [8]. The window size on this data set is 3 years: each collaboration is instantaneous in the original data sets, but Arnetminer expands these collaborations to exist for three years past this. The resolution is a single year, with data for every year from 1986 up to and including 2012. The growth of a particularly successful researcher’s collaboration network is shown in Fig. 3.

The data was cleaned in order to consider researchers for whom we have a full picture of their careers. Any researchers who were active in the first year of the data set were removed to exclude researchers who started their careers before the data began, and any researchers still active at the end of the data set, with the assumption that these researchers may not have

finished their careers yet, were also removed. This eliminated approximately 50% of the data set. 10000 authors were then sampled to reduce the computational complexity. Finally, any authors that did not have careers of at least 5 years were discarded as this was found to make prediction easier with stronger correlations, leaving us with about 2000 authors. The lifetime distribution is shown in Fig. 1.

We postulate the following:

Hypothesis 1: Some components of graphlet profiles of researchers’ co-authorship ego-networks at an early stage of their careers correlate with their career lengths, after controlling for potentially confounding variables.

Hypothesis 2: Graphlet profiles coloured to distinguish egos from the alters correlate more strongly with career length.

B. Correlation of features with lifetime

The correlation of each graphlet with career longevity is shown in Fig. 4. There are no strong correlations, but we can see that 4.5c, 3.2c, 4.3b, and 4.6b have weak positive correlations. If one cross-references with Fig. 2, it can be seen that these are the graphlets where the ego is central and of higher degree than its neighbours. Contrary to what one might expect, 4.2b and 3.1b, the all-to-all graphlets, have a very weak negative correlation with lifetime.

On the whole, the uncoloured treatment has weaker correlations with lifetime and misses out on some subtlety: where monochrome graphlets have equivalent coloured graphlets that are both negatively and positively correlated with lifetime, the net effect is to have an overall correlation which is closer to zero, which could be misleading. This is especially evident with graphlet 4.3: 4.3b is one of the strongest positive predictors of lifetime, but 4.3a is negatively correlated in the coloured treatment, albeit not significantly so. This leads to 4.3 having no overall correlation with lifetime in the uncoloured treatment, which could explain why previous work found this graphlet to be unremarkable [10].

The higher correlation with lifetime for 3.2c and 4.3b could be because they are proxies for the number of papers published; 4.5c could be a proxy for the number of papers published by your collaborators. The fact that the correlation is stronger than that of degree centrality could be because degree centrality is a poor measure of the number of papers published in a co-authorship network; the negative correlation of the clustering co-efficients is further evidence in favour of this as a higher number of authors per paper will result in a higher clustering co-efficient.

To investigate whether graphlet profiles add any new insight compared to the standard node centralities, pairwise correlation analysis on the graphlet profile components and the centralities was performed, as shown in Fig. 5. It can be seen that the monochromatic motifs 4.5a, 4.3a, 4.6a, 3.2a and 2.1a are closely related, which is as expected as they are all subgraphs of 4.6a. Of particular interest is the graphlets that correlate with lifetime also correlate with the various centralities: 4.3b, 4.5c, 4.6b and 3.2c all correlate strongly with betweenness centrality. Generally, this correlation plot

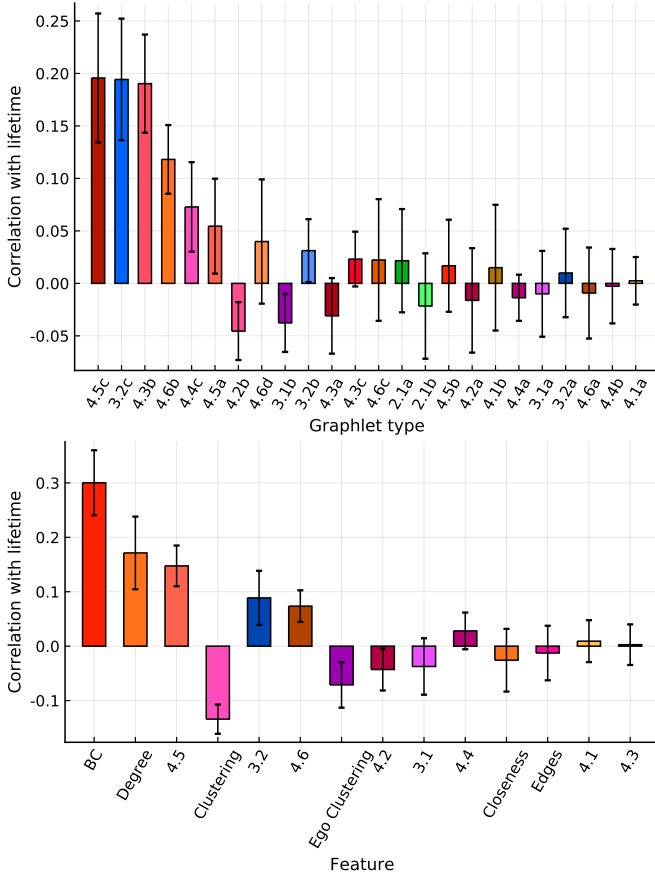


Fig. 4. The correlation of each graphlet frequency with lifetime for graphlets in the coloured treatment (top) and uncoloured treatment (bottom).

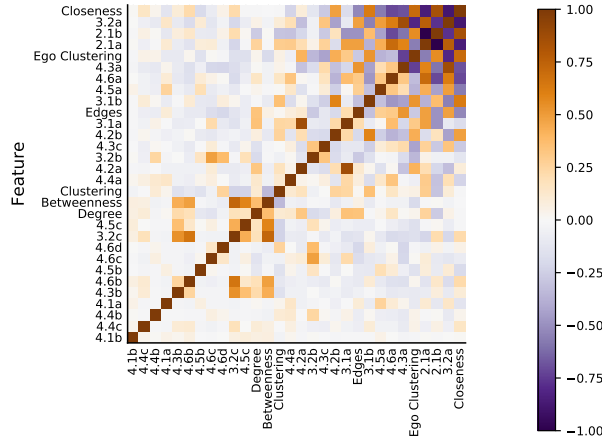


Fig. 5. A pairwise correlation matrix of the various features on the subsample of co-authorship 3-ego-networks.

was unsurprising, and confirmed the need for further analysis to rule out degeneracies.

C. Prediction of researcher lifetimes

Due to the high pairwise correlations between some of the centralities and the graphlets that were most highly correlated with lifetime, we tested for degeneracy amongst the features

TABLE II
REGRESSION ANALYSIS FOR GRAPHLETS IN THE MONOCHROMATIC TREATMENT VERSUS RESEARCH LIFETIME

Feature	β	Std.Error	t value	Pr(> t)
3.1a	0.4	0.45	0.9	0.3688
4.3a	1.8	0.78	2.3	0.0227
4.5a	2.9	0.92	3.1	0.0018
Betweenness	3.9	0.5	7.7	$< 1e - 13$
Clustering	-0.69	0.22	-3.2	0.0016
Degree	0.1	0.024	4.4	$< 1e - 4$
Edges	-0.0026	0.00086	-3.0	0.0027

The features are described in detail in Table I; for the meanings of the other columns see Sect. III-E.

TABLE III
REGRESSION ANALYSIS FOR COLOURED GRAPHLETS VERSUS RESEARCH LIFETIME

Feature	β	β_{reduced}	Std.Error	t value	Pr(> t)
3.1a	1.8	10.0	6.9	1.5	0.1385
3.1b	0.95	4.0	1.6	2.6	0.0095
4.3a	-4.1	-0.95	3.4	-1.2	0.2280
4.3b	10.0	0.34	2.7	3.8	0.0001
4.3c	1.3	0.077	1.8	0.69	0.4882
Betweenness	4.2	0.74	0.75	5.7	$< 1e - 7$
Clustering	-0.82	-0.00098	0.23	-3.5	0.0005
Degree	0.12	-0.0084	0.028	4.3	$< 1e - 4$

The features are described in detail in Table I; for the meanings of the other columns see Sect. III-E.

by performing linear regression analysis. The results can be seen in Table II and Table III. Graphlet profile components with particularly insignificant results have been omitted for brevity. Many of the most strongly correlated graphlets with lifetime, 4.5c, 3.2c, and 4.6b, were found to be degenerate to betweenness centrality. It can be seen that many of the centralities are much more significant than the graphlet profile components.

Comparing Table II with Table III, we can accept hypothesis 2: on the whole, the coloured graphlet profiles have more significant and stronger relationships with lifetime than in the monochrome treatment. We can also accept hypothesis 1: graphlet 4.3b is strongly and significantly linked with career lifetime, even after accounting for degree and betweenness centrality, which which it is correlated. It is displayed in Fig. 6. The large co-efficient suggests that the graphlet's presence could explain a large amount of research lifetime, which could warrant further investigation. There are no other graphlets which are significantly linked to lifetime.

Some of the centralities are highly significant; betweenness centrality in particular has a very significant link, although it should be noted that it can only explain 4 years of lifetime. The significance of betweenness centrality contradicts earlier research, although the significant but small influence of degree on lifetime is in agreement [16].

Curiously, graphlets 4.1a and 4.1b do not appear at all in our data set, despite them being previously found to be important for success as measured by citation count [10]. This distinction



Fig. 6. Graphlet 4.3b from Fig. 4, the only graphlet found to have a significant link with career lifetime, after controlling for many standard node centralities.

could be because that previous work looked at motifs (i.e. all subgraphs, not just induced ones), but one would expect to see the 4.2 and 4.4 graphlets, of which 4.1 is a subgraph, correlating positively with lifetime with some certainty, which is not the case. 4.2 seems to be negatively correlated with lifetime, if at all.

Finally, it is pertinent to note that none of the features tested have a very large impact on career-length, especially when considering the reduced co-efficients, suggesting that there is a large element of random chance, or that some important factor has been left out of our analysis, which is consistent with other work [5].

V. CONCLUSION

We have demonstrated that coloured graphlet profiles can be predictors for a real-world characteristic of nodes, and this predictive power is not explained by the other standard centralities; particularly, degree centrality and betweenness centrality.

In our application, the 4.3b graphlet, the 4-star centred on the author of interest, was found to be the strongest predictor of career length in researchers in the fifth year of their career. Therefore, the best advice we can give to young researchers is that they should try to collaborate broadly with people who do not collaborate with each other. However, the correlations we have found do come with the usual caveat: there could be some other confounding variables that are at play—for example, it could be that the kind of people who collaborate with lots of people are interdisciplinarians who are particularly intelligent and therefore were more likely to be successful anyway, or that the success of the 4.3b graphlet compared to degree centrality could be because 4.3b requires more papers to be published than degree centrality; indeed, other work has found that the number of papers published is one of the main factors in determining career-length [5]. This could be solved by a more thorough study with more information about each author. It could also be beneficial to use a data set spanning more time as our current approach limits the maximum career length to 23 years—we disregard any researcher with a career longer than that.

Finally, it would be interesting to see if our technique has similar or greater success in other applications. Analysis of any network where there are interesting properties on the nodes which are linked to the medium scale structure of the network could plausibly benefit; for example, it could be used to detect groups of users on a social network spreading information in an inauthentic fashion.

Our main conclusion is that when characterising a node through its ego-network, it can be useful to retain the information as to which node is the ego, and which are the alters. This information can be used by existing graph packages by colouring the ego one colour and the alters another.

REFERENCES

- [1] S. H. Strogatz, “Exploring complex networks,” *Nature*, vol. 410, no. 6825, pp. 268–276, Mar. 2001.
- [2] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, “Network motifs: Simple building blocks of complex networks,” *Science*, vol. 298, no. 5594, pp. 824–827, Oct. 2002.
- [3] N. Pržulj, “Biological network comparison using graphlet degree distribution,” in *Bioinformatics*, 2007.
- [4] P. Cunningham, M. Harrigan, G. Wu, and D. O’Callaghan, “Characterizing ego-networks using motifs,” *Network Science*, vol. 1, no. 2, pp. 170–190, 2013.
- [5] S. Milojević, F. Radicchi, and J. P. Walsh, “Changing demographics of scientific careers: The rise of the temporary workforce,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 50, p. 201800478, 2018.
- [6] M. Biryukov and C. Dong, “Analysis of computer science communities based on DBLP,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6273 LNCS, pp. 228–235, 2010.
- [7] S. Priddy, “LSBF Careers Report: Are UK professionals looking to change careers?” London School of Business & Finance, London, Tech. Rep., 2015.
- [8] H. Zhuang, Y. Sun, J. Tang, J. Zhang, and X. Sun, “Influence Maximization in Dynamic Social Networks,” in *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, 2013, pp. 1313–1318.
- [9] D. O’Callaghan, M. Harrigan, J. Carthy, and P. Cunningham, “Network Analysis of Recurring YouTube Spam Campaigns,” pp. 531–534, 2012.
- [10] L. Krumov, C. Fretter, M. Müller-Hannemann, K. Weihe, and M. T. Hütt, “Motifs in co-authorship networks and their relation to the impact of scientific publications,” *European Physical Journal B*, vol. 84, no. 4, pp. 535–540, 2011.
- [11] A. Masoudi-Nejad, F. Schreiber, and Z. Kashani, “Building blocks of biological networks: a review on major network motif discovery algorithms,” *IET Systems Biology*, vol. 6, no. 5, pp. 164–174, 2012.
- [12] Z. Kashani, H. Ahrabian, E. Elahi, A. Nowzari-Dalini, E. Ansari, S. Asadi, S. Mohammadi, F. Schreiber, and A. Masoudi-Nejad, “Kavosh: a new algorithm for finding network motifs,” *BMC Bioinformatics*, vol. 10, no. 1, p. 318, Oct. 2009.
- [13] S. Kumar, “Co-authorship networks: A review of the literature,” *Aslib Journal of Information Management*, vol. 67, no. 1, pp. 55–73, 2015.
- [14] S. Choobdar, P. Ribeiro, S. Bugla, and F. Silva, “Comparison of Co-authorship Networks across Scientific Fields Using Motifs,” *2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 147–152, 2012.
- [15] T. Chakraborty, N. Ganguly, and A. Mukherjee, “An author is known by the context she keeps: significance of network motifs in scientific collaborations,” *Social Network Analysis and Mining*, vol. 5, no. 1, pp. 1–21, 2015.
- [16] E. Y. Li, C. H. Liao, and H. R. Yen, “Co-authorship networks and research impact: A social capital perspective,” *Research Policy*, vol. 42, no. 9, pp. 1515–1530, 2013.
- [17] R. S. Burt, “Models of Network Structure,” *Annual Review of Sociology*, 1980.
- [18] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, “Julia: A Fresh Approach to Numerical Computing,” *SIAM Review*, vol. 59, no. 1, pp. 65–98, Jan. 2014.
- [19] S. Bromberger, “JuliaGraphs/LightGraphs.jl: LightGraphs v0.10.5,” Sep. 2017.
- [20] B. D. McKay and A. Piperno, “Practical graph isomorphism, II,” *Journal of Symbolic Computation*, vol. 60, pp. 94–112, Jan. 2014.
- [21] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to linear regression analysis*. John Wiley & Sons, 2012, vol. 821.
- [22] C. Bonferroni, “Sulle medie multiple di potenze,” *Bollettino dell’Unione Matematica Italiana*, vol. 5, no. 3–4, pp. 267–270, 1950.